Bandit Multiclass Linear Classification: Efficient Algorithms for the Separable Case

YAHOO! RESEARCH — Microsoft Research

Alina Beygelzimer Dávid Pál Balázs Szörényi (Yahoo! Research, New York City)
Devanathan Thiruvenkatachari (New York University)
Chen-Yu Wei (University of Southern California)
Chicheng Zhang (Microsoft Research, New York City)

NYU — University of Southern California

Abstract

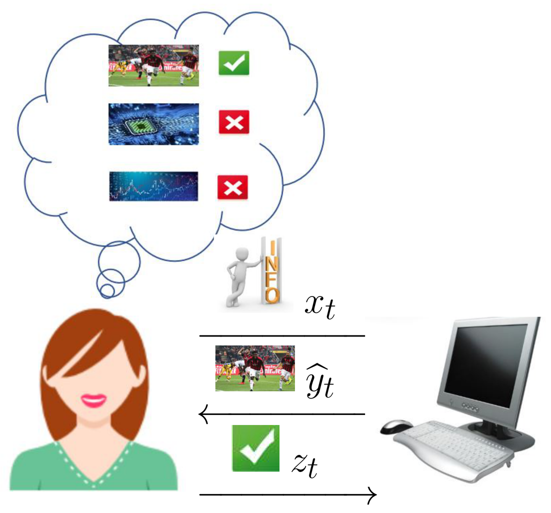
We design efficient algorithms for online bandit K-class linear classification when the data is linearly separable by a margin γ. We consider two notions of linear separability, *strong* and *weak*.

1. Under the strong linear separability condition, we design an efficient algorithm that achieves a near-optimal mistake bound $\tilde{O}(\frac{K}{\gamma^2})$.
2. Under the more challenging weak linear separability condition, we design an efficient algorithm with a mistake bound quasi-polynomial in $\frac{1}{\gamma}$ for constant K. Our key observation is a reduction from the weak linear separability to strong linear separability condition via a specialized nonlinear mapping.

Online Bandit Linear Classification

For $t = 1, 2, \ldots, T$:

1. Example (x_t, y_t) is chosen, where
   $x_t \in \mathbb{R}^d$ is the feature (shown to the learner),
   $y_t \in [K]$ is the label (hidden).
2. Predict class label $\hat{y}_t \in [K]$.
3. Observe feedback $z_t = \mathbb{1}[\hat{y}_t \neq y_t] \in \{0, 1\}$.



Goal: minimize the total number of mistakes $\sum_{t=1}^{T} z_t$.

Technical assumption: $\|x_t\| \leq 1$ for all t.

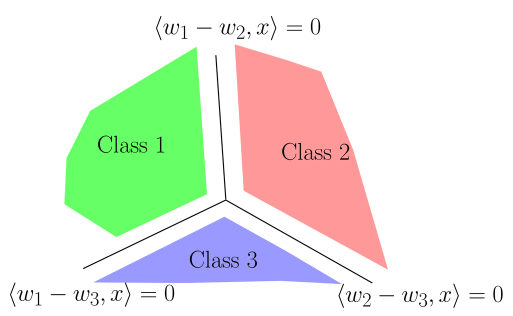
Notions of Linear Separability

Multiclass linear classification: classifier $W = (w_1, w_2, \ldots, w_K) \in \mathbb{R}^{K \times d}$ predicts on x by:

1. Compute i-th score $\langle w_i, x \rangle$ for each label i
2. Predict $\hat{y} = \text{argmax}_i \langle w_i, x \rangle$

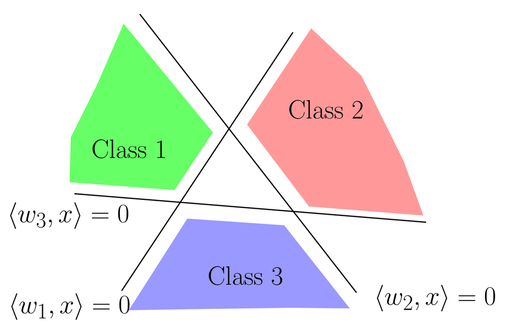
Weakly linearly separable: there exists W^* with $\|W^*\|_F \leq 1$, and for all (x, y):

$$\forall y' \neq y, \quad \left\langle w^*_y, x \right\rangle \geq \left\langle w^*_{y'}, x \right\rangle + \gamma \,,$$



Strongly linearly separable: there exists W^* with $\|W^*\|_F \leq 1$, and for all (x, y):

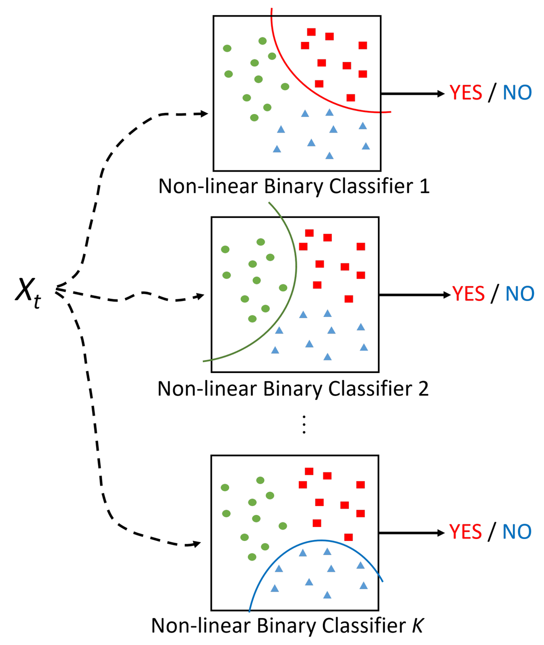
$$\left\langle w^*_y, x \right\rangle \geq \gamma/2 \,,$$
$$\forall y' \neq y, \quad \left\langle w^*_{y'}, x \right\rangle \leq -\gamma/2 \,.$$



Algorithm

Key idea:

1. Create K online classification tasks $T_i, i = 1, \ldots, K$, where task T_i is to predict whether examples belong to class i.
2. For each task T_i, maintain a separate online classification algorithm $\mathcal{A}_i$.
3. When predicting, aggregate the predictions from all $\mathcal{A}_i$'s; after receiving the feedback, update all $\mathcal{A}_i$'s .



for $t = 1, 2, \ldots, T$: **do**

Receive example x_t.

Query: For $i = 1, \ldots, K$, ask algorithm $\mathcal{A}_i$ whether x_t belongs to class i.

Predict:
Case 1: If ≥ 1 of them respond YES:
$\hat{y}_t \leftarrow$ any one of those YES labels

Case 2: If all of them respond NO:
$\hat{y}_t \leftarrow$ uniform from $\{1, \ldots, K\}$

Receive feedback $z_t = \mathbb{1}[\hat{y}_t \neq y_t]$.

Update:
Case 1: If $z_t = 1$, send example (x_t, NO) to $\mathcal{A}_{\hat{y}_t}$.
Case 2: If $z_t = 0$, send example (x_t, YES) to $\mathcal{A}_{\hat{y}_t}$.

Theorem 1. *If for each i, $\mathcal{A}_i$ makes at most M_i mistakes for task T_i, then our proposed algorithm makes at most $K(M_1 + \ldots + M_k)$ in expectation.*

Performance Guarantees

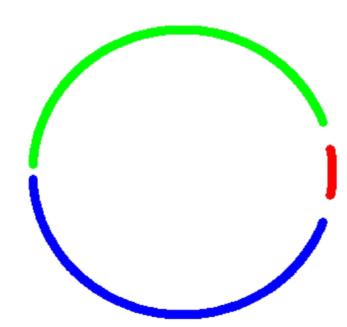
Setting	Sublearner $\mathcal{A}$	Sublearner mistake bound	Mistake bound
Strongly linearly separable	Perceptron	$O(1/\gamma^2)$	$O(K/\gamma^2)$ (tight)
Weakly linearly separable	kernel Perceptron with rational kernel $k(x, x') = \frac{1}{1 - \frac{1}{2}\langle x, x'\rangle}$	$2^{\tilde{O}(\min(K \log^2(1/\gamma), \sqrt{1/\gamma} \log K))}$	$2^{\tilde{O}(\min(K \log^2(1/\gamma), \sqrt{1/\gamma} \log K))}$

Related Work - Weakly Linearly Separable Setting

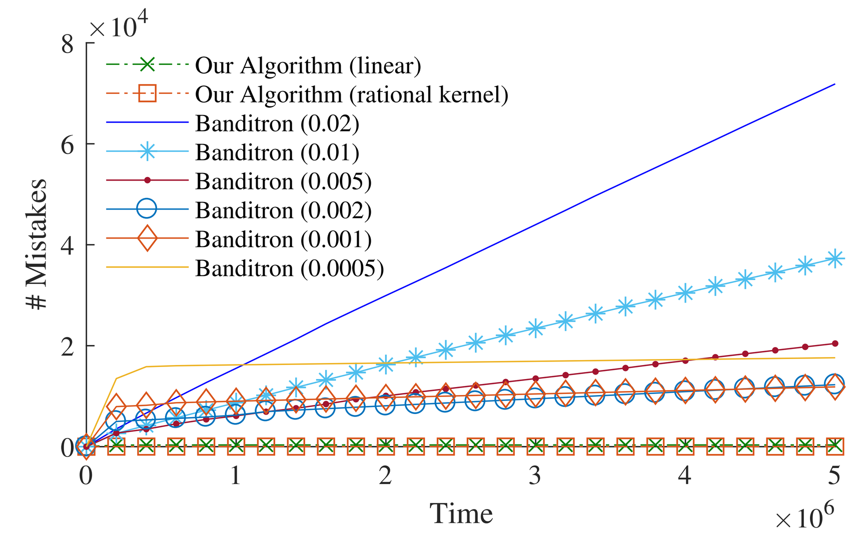
Algorithm	Mistake bound (in Big-O)	Efficient?
Halving [Kakade et al., 2008]	$K^2(\ln T)/\gamma^2$ or $dK^2 \ln(1/\gamma)$	No
Minimax algorithm [Daniely and Helbertal, 2013]	$\min(K/\gamma^2, dK \ln(1/\gamma))$ (tight)	No
Banditron [Kakade et al., 2008], Newtron [Hazan and Kale, 2011], SOBA [Beygelzimer et al., 2017], OBAMA [Foster et al., 2018]	at least $\sqrt{KT/\gamma^2}$ or $\sqrt{dKT}$	Yes

Empirical Evaluation

Experiment 1: strongly separable setting. Our algorithm with linear Perceptron and rational kernel Perceptron performs well and exhibit finite mistake bound experimentally.

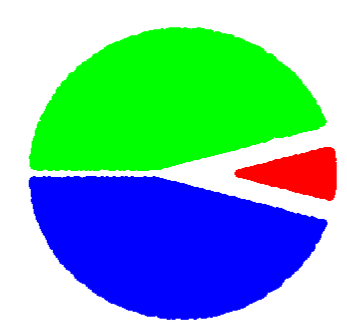


(a) A strongly separable data distribution with $\gamma = 0.05$.

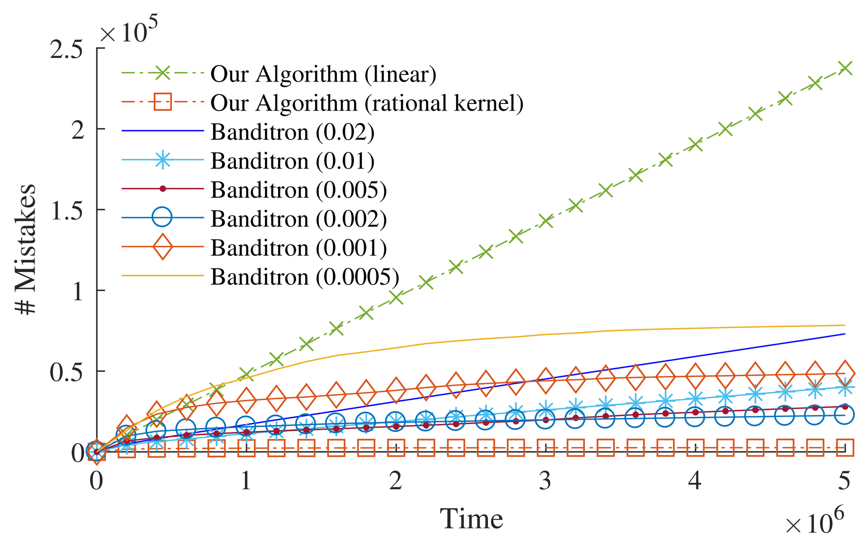


(b) Cumulative number of mistakes as a function of number of examples seen.

Experiment 2: weakly separable setting. Our algorithm with rational kernel Perceptron performs well and exhibit finite mistake bound experimentally. Our algorithm with linear Perceptron has a high number of mistakes, which is within expectation.



(a) A weakly separable data distribution with $\gamma = 0.05$.



(b) Cumulative number of mistakes as a function of number of examples seen.

Hardness Results

1. Any "ignorant algorithm" will make $\Omega(\min\{\sqrt{T}, 2^{\Omega(d)}\})$ mistakes even when the data is strongly linearly separable. An ignorant algorithm does not update itself when it makes a mistake (variants of SOBA [Beygelzimer et al., 2017] and OBAMA [Foster et al., 2018] are of this type).
2. Finding a linear classifier that agrees with a labeled dataset and a complementary labeled dataset is NP-hard (naive algorithm requires $2^{O(d)}$ computational complexity).
   A complementary labeled dataset [Ishida et al., 2017] consists of the following types of examples:

   $$(x, \overline{y}) : \text{example } x \text{ does not belong to class } y$$

   Observation: finding an intersection of two halfspaces that agrees with a dataset is NP-hard [Blum and Rivest, 1993].